Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection

Christophe Biernacki, Alexandre Lourme

Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection

Christophe Biernacki, Alexandre Lourme

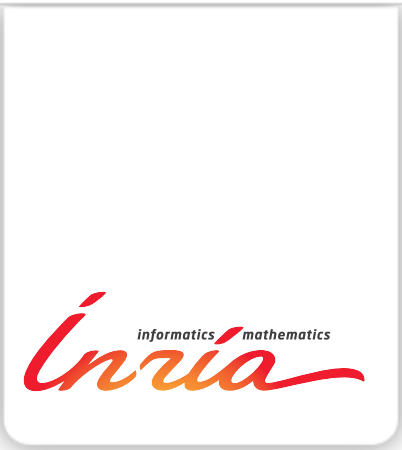
RESEARCH REPORT

N° 7932

April 2012

Project-Team Modal

ISSN 0249-6399 ISRN INRIA/RR--7932--FR+ENG



Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection

Christophe Biernacki*, Alexandre Lourme†

Project-Team Modal

Research Report n° 7932 — April 2012 — 21 pages

Abstract: Gaussian mixture model-based clustering is now a standard tool to determine an hypothetical underlying structure into continuous data. However many usual parsimonious models, despite their appealing geometrical interpretation, suffer from major drawbacks as scale dependence or unsustainability of the constraints by projection. In this work we present a new family of parsimonious Gaussian models based on a variance-correlation decomposition of the covariance matrices. These new models are stable by projection into the canonical planes and, so, faithfully representable in low dimension. They are also stable by modification of the measurement units of the data and such a modification does not change the model selection based on likelihood criteria. We highlight all these stability properties by a specific geometrical representation of each model. A detailed GEM algorithm is also provided for every model inference. Then, on biological and geological data, we compare our stable models to standard geometrical ones.

Key-words: Correlation, EM algorithm, Faithful projection, Maximum-Likelihood, Standard deviation, Unit independence

* University Lille 1 & CNRS & Inria
† IUT département de Génie Biologique, Université de Pau et des Pays de l'Adour

RESEARCH CENTRE
LILLE – NORD EUROPE

Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq

Classification par modèles gaussiens parcimonieux, invariants aux unités et stables par projections

Résumé : La classification à base de modèles de mélanges gaussiens est maintenant un outil standard pour déterminer une hypothétique structure cachée dans un jeu de données continues. Pourtant de nombreux modèles parcimonieux usuels, malgré leur interprétation géométrique conviviale, souffrent de défauts majeurs comme la dépendance aux unités de mesure ou encore la violation des contraintes par projection. Dans ce travail, nous présentons une nouvelle famille de modèles gaussiens parcimonieux reposant sur une décomposition variance-corrélation des matrices de covariance. Ces nouveaux modèles sont stables par projection sur les plans canoniques et, par conséquent, fidèlement représentables en faible dimension. Ils sont aussi indépendants des unités de mesure des données, ce qui signifie que ce choix parfois arbitraire n'a aucune conséquence sur la sélection de modèle reposant sur des critères à base de vraisemblance. Nous mettons en évidence toutes ces propriétés de stabilité par une représentation géométrique spécifique à chacun des modèles. Un algorithme GEM est aussi donné en détail pour estimer leurs paramètres. Nous comparons enfin nos modèles stables et les modèles géométriques standards sur des données réelles issues de la biologie et de la géologie.

Mots-clés : algorithme EM, corrélation, écart-type, indépendance aux unités, fidélité de projection, maximum de vraisemblance

1 INTRODUCTION

Nowadays Gaussian mixture models are commonly used for classifying continuous data. They allow both (i) to unambiguously determine the structure of a dataset by defining rigorously the concept of homogeneous subgroups and (ii) to provide a meaningful interpretation of the inferred partition. In order to reduce gradually the variability of the general heteroscedastic model, Celeux and Govaert (1995), inspired by Banfield and Raftery (1993), define some geometrical parsimonious Gaussian mixtures based on a spectral decomposition of the covariance matrices. These models have had a seminal influence in recent years (see Biernacki 1997; Biernacki, Celeux, Govaert, and Langrognet 2006; Bouveyron 2006; Baudry 2009; Greselin, Ingrassia, and Punzo 2011) and nowadays they are very widespread. They enable Bouveyron, Girard, and Schmid (2007) for example, to detect classes into the chemical composition of Mars soil. They are employed by Michel (2008) to classify production curves and to determine the nature of oil fields. They are used also by Maugis et al. (2009) for selecting variables intended to clarify the gene functions.

However some of these geometrical models suffer from multiple drawbacks. Projecting a model onto a canonical subspace for example, may break the model structure. Then some of the geometric models cannot be represented faithfully in low dimension. In addition the geometric models are not stable by changing the measurement units: such a modification may infringe again the model structure. Another consequence is that the model selected within the geometric family thanks to a classical likelihood criterion like AIC (Akaike 1974) or BIC (Schwarz 1978) depends on the measurement units. Thus the retained model does not really reflect some intrinsic property of the data.

We display in this work a new family of parsimonious Gaussian mixtures based on a variance-correlation decomposition of the covariance matrices. The parsimony of our models refers to parameters of statistical interpretation (standard deviation, correlation, coefficient of variation) instead of a geometric interpretation (volume, orientation, shape). They own multiple stability properties which make them mathematically consistent and facilitate their interpretation. Firstly, the characteristic constraints of each model still remain in every canonical plane. This ensures that each parsimonious mixture can be represented faithfully in dimension 2. Secondly, changing the measurement units does not alter the constraints inherent to the models. In addition the choice of some particular units does not even have any effect on the model selection based on many classical criteria. Especially raw data and reduced data lead to select the same model.

We remind in Subsection 2.1 the general framework of the Gaussian mixture model-based clustering method and then, in Subsection 2.2, what are the standard geometric models of Celeux and Govaert (1995). Then we define our new Gaussian mixtures based on a statistical interpretation of the classes (Section 3); a geometrical representation of them is proposed at the same time. Section 4 highlights the stability properties of our new model family which are lacking in the geometric family of Celeux and Govaert (1995). We show in Subsection 4.1 that any mixture of this family can be faithfully represented in any canonical plane. Then we establish in Subsection 4.2 that our models are stable by changing the measurement units and that such a modification has no effect on the model selection when the latter is based on classical likelihood criteria like AIC (Akaike 1974), BIC (Schwarz 1978) or ICL (Biernacki, Celeux, and Govaert 2000). Within this model family, the Maximum Likelihood parameter estimation relies on a GEM algorithm which is detailed in Section 5. In Section 6 we compare on real data our models with the standard geometrical ones. First, on a very famous dataset concerning eruptions of the Old

Faithful geyser, we illustrate (Subsection 6.1) the scale invariance (resp. the scale dependence) of the model selection within the new family (resp. within the geometric family). In this geological context we will see that the new models both (i) improve the fit of the geometrical models and (ii) lead to a more convincing interpretation of the properties of the conditional data. Then in Subsection 6.2 the new models are used in order to classify a sample of seabirds described by morphological features. The new family enables to retrieve the bird subspecies better than the geometrical family does; moreover the selected new model allows to interpret the bird subspecies as arising stochastically from some common reference population. At last we evoque in Section 7 some results from additional experiments and we consider several perspectives of our new Gaussian mixtures.

2 GEOMETRICAL PARSIMONIOUS MODELS

2.1 General model-based clustering principle

Unsupervised classification aims to (i) decide if the data within some sample $\mathbf{x} = \{\mathbf{x}_i; i = 1, \ldots, n\} \subset \mathbb{R}^d$ are homogeneous and otherwise (ii) to detect some underlying partition into $\mathbf{x}$. So a matrix $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$ has to be determined where the i^{th} row $\mathbf{z}_i = (z_i^1, \ldots, z_i^K)$ indicates whether $\mathbf{x}_i$ belongs to the class k ($z_i^k = 1$) or not ($z_i^k = 0$). $K \in \mathbb{N}^*$ represents the (unknown) cluster number.

Gaussian model-based clustering assumes that the couples $(\mathbf{x}_i, \mathbf{z}_i)$ are realizations of independent random vectors identically distributed to $(\mathbf{X}, \mathbf{Z})$. The k^{th} component Z^k of $\mathbf{Z} \in \{0,1\}^K$ equals 1 (and the other ones 0) with probability π_k ($0 < \pi_k < 1$ and $\sum_{k=1}^K \pi_k = 1$) and the conditional vector $(\mathbf{X}|Z^k = 1)$ is normal, non-degenerate, with center $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and with covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$. So the observed data $\mathbf{x}_i$ are assumed to be distributed according to the Gaussian mixture:

$$f(\bullet; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \Phi_d(\bullet; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

$\Phi_d(\bullet; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denoting the normal density of center $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\psi} = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k); k = 1, \ldots, K\}$ denoting the parameter of the model. In addition the missing data $\mathbf{z}_i$ are assumed to be distributed according to the K-dimensional multinomial distribution of order 1 and parameter $(\pi_1, \ldots, \pi_K)$.

Noting $\hat{\boldsymbol{\psi}}$ the Maximum Likelihood estimate of $\boldsymbol{\psi}$, then data are classified by Maximum A Posteriori (MAP): $\hat{z}_i^k = 1 \Leftrightarrow \forall j \in \{1, \ldots, K\},\ t_i^k \geq t_i^j$, where t_i^k is the conditional probability

$$t_i^k = \hat{\pi}_k \Phi_d(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k) / f(\mathbf{x}_i; \hat{\boldsymbol{\psi}}). \tag{2}$$

So clustering based on Gaussian mixtures consists of two steps: (i) the inference of a model from the observed data $\mathbf{x}_i$ and then (ii) the assessment of classes by estimating the missing data $\mathbf{z}_i$.

The step (i) is an opportunity to make compete several parsimonious hypotheses that is to consider diverse restrictions of the parameter space $\boldsymbol{\Psi}$. This step enables also to propose several values of the mixture order K. The BIC criterion (Schwarz 1978) enables to choose both a parsimonious mixture model and a cluster number K. This criterion is defined by:

$$BIC = (\eta/2) \log n - \ell(\hat{\boldsymbol{\psi}}; \mathbf{x}), \tag{3}$$

where η denotes the dimension of $\boldsymbol{\psi}$ parameter and $\ell(\hat{\boldsymbol{\psi}}; \mathbf{x})$ its maximized log-likelihood computed on $\mathbf{x}$. As BIC leads sometimes to strongly overlapping groups which are difficult to interpret, one may prefer $ICL = BIC - \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_i^k \log t_i^k$ (see Biernacki, Celeux, and Govaert 2000). Indeed this other likelihood-based criterion favours well separated groups and more interpretable structures.

2.2 Spectral decomposition

As the Gaussian components are non-degenerate, each covariance matrix $\boldsymbol{\Sigma}_k$ is symmetric, definite, positive. Then $\boldsymbol{\Sigma}_k$ can be decomposed as:

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{S}_k \boldsymbol{\Lambda}_k \mathbf{S}_k', \tag{4}$$

where: (i) $\lambda_k = |\boldsymbol{\Sigma}_k|^{1/d}$ (volume of the class k), (ii) $\mathbf{S}_k$ is an orthogonal matrix the columns of which are $\boldsymbol{\Sigma}_k$ eigenvectors (orientation of the class k) and (iii) $\boldsymbol{\Lambda}_k$ is a diagonal definite positive matrix with determinant 1 and with diagonal coefficients in decreasing order (shape of the class k).

A Gaussian mixture of Celeux and Govaert (1995) is a combination of parsimonious hypotheses on λ_k, $\mathbf{S}_k$ and $\boldsymbol{\Lambda}_k$ parameters. For example the so-denoted $[\lambda \mathbf{S}_k \boldsymbol{\Lambda} \mathbf{S}_k']$ geometric model (illustrated by Figure 1 in case $K = 2$) assumes that the Gaussian components have identical shapes, same volumes and free orientations (this model is called homometroscedastic in Greselin et al. 2011). But the constraints of this model do not remain in the canonical subspaces, as shown by Figure 1.

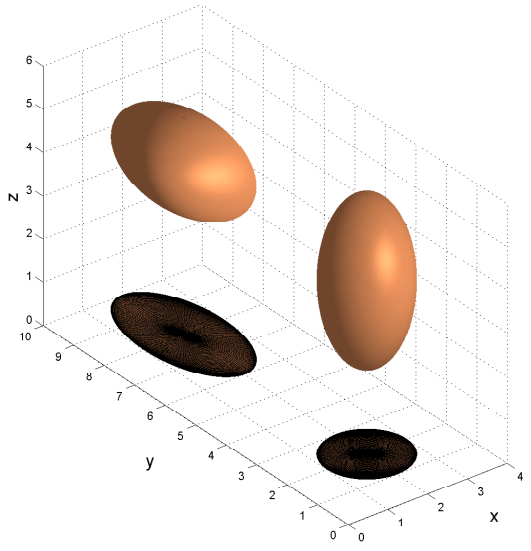
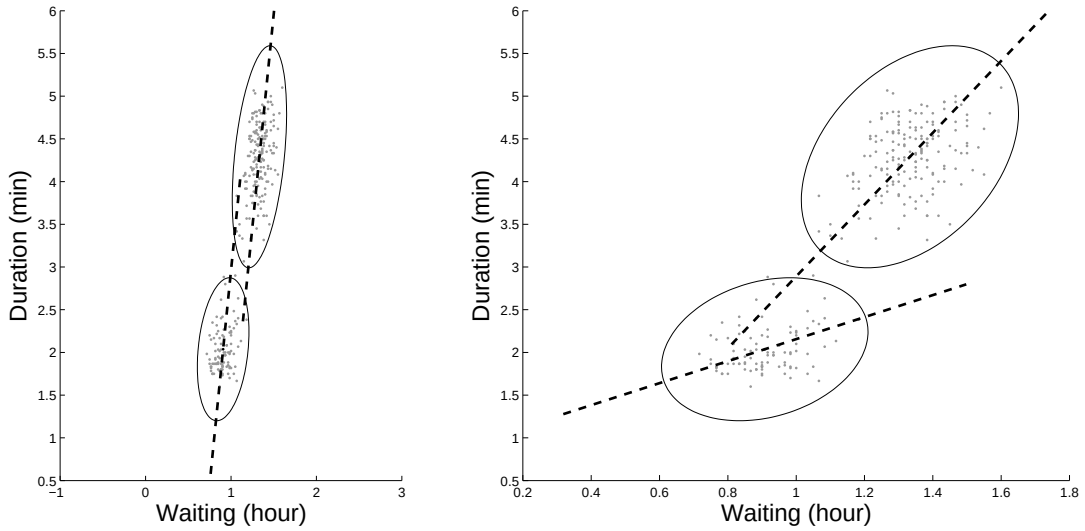


Figure 1: *A major drawback of the geometrical models: unsustainability of the structure by projection into the canonical planes.*

For another example, $[\lambda_k \mathbf{S} \mathbf{\Lambda}_k \mathbf{S}']$ assumes that the orientations of the classes are homogeneous whereas the volumes and the shapes are free (this model is called homotroposcedastic in Greselin et al. 2011). Figure 2a represents in an orthonormal basis two Gaussian components inferred under these assumptions on the famous Old Faithful data (described in Subsection 6.1). But Figure 2b shows that a non-isotropic axis rescaling infringes the hypothesis of homogeneously orientated classes. This illustrates a second drawback of the geometrical models: they are not scale invariant.



(a) *Gaussian classes with same orientation in an orthonormal basis.*

(b) *A modification of x-axis scale infringes the assumption of homogeneous orientations.*

Figure 2: *Another drawback of the geometrical models: unsustainability of the structure by non-isotropic axis rescaling.*

3 NEW PARSIMONIOUS MODELS

3.1 Variance-correlation decomposition

As they are symmetric, definite, positive, the covariance matrices can be also decomposed as:

$$\mathbf{\Sigma}_k = \mathbf{T}_k \mathbf{R}_k \mathbf{T}_k \tag{5}$$

where $\mathbf{T}_k$ is the corresponding diagonal matrix of conditional standard deviations and $\mathbf{R}_k$ the associated matrix of conditional correlations. So $\mathbf{T}_k(i,j) = \sqrt{\mathbf{\Sigma}_k(i,j)}$ if $i = j$ and 0 otherwise, and $\mathbf{R}_k = (\mathbf{T}_k)^{-1} \mathbf{\Sigma}_k (\mathbf{T}_k)^{-1}$. Contrarily to many other decompositions as Cholesky's, (5) is canonical since both $\mathbf{T}_k$ and $\mathbf{R}_k$ matrices are unique.

The decomposition (5) allows to consider several models by combining meaningful constraints on $\mathbf{T}_k$ and $\mathbf{R}_k$ parameters but on $\boldsymbol{\mu}_k$ centers as well:

- $\mathbf{T}_k$ $(k = 1, \ldots, K)$ matrices are diagonal definite positive. We consider three possible states of standard deviations: free (no additional constraint on $\mathbf{T}_k$ matrices), isotropically transformed $(\forall\ (k,k') : \mathbf{T}_{k'} = a_{k,k'} \mathbf{T}_k ; a_{k,k'} \in \mathbb{R}_+^*)$ or homogeneous $(\mathbf{T}_k = \mathbf{T})$.

- $\mathbf{R}_k$ $(k = 1, \ldots, K)$ matrices are symmetric definite positive and their diagonal coefficients equal 1. We consider two possible states of the correlations: free (no additional constraint on $\mathbf{R}_k$ matrices) or homogeneous ($\mathbf{R}_k = \mathbf{R}$).

- Vectors $\mathbf{V}_k = \mathbf{T}_k^{-1}\boldsymbol{\mu}_k$ $(k = 1, \ldots, K)$–the components of which are conditional first-order-standardized-moments–are free or equal ($\mathbf{V}_k = \mathbf{V}$). When $\boldsymbol{\mu}_k$ components are non-zero, the inverses of $\mathbf{V}_k$ components are conditional coefficients of variation. So $\mathbf{V}_k = \mathbf{V}$ means also that the conditional coefficients of variation are supposed to be homogeneous.

The so-called RTV family consists of eleven Gaussian mixture models obtained by combining the previous constraints on the conditional correlations, standard deviations and first-order-standardized-moments. The family does not include the model assuming all parameters $\mathbf{T}_k$, $\mathbf{R}_k$ and $\mathbf{V}_k$ as homogeneous because this combination amounts to merge all the components of the mixture.

Let us note two meaningful differences between $\mathbf{T}_k$ and $\mathbf{R}_k$ parameters. Firstly a constraint on $\mathbf{T}_k$ matrices postulates a model intrinsic to each variable whereas a constraint on $\mathbf{R}_k$ matrices involves a model on couples of variables. Secondly $(\mathbf{V}_k, \mathbf{R}_k)$ is a Gaussian parameter obtained by normalizing $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ thanks to $\mathbf{T}_k$. Indeed the normal vector of reduced variables $(\mathbf{T}_k)^{-1}(\mathbf{X}|Z^k = 1)$ has center $\mathbf{V}_k$ and covariance matrix $\mathbf{R}_k$.

The most general RTV model assumes $\mathbf{R}_k$, $\mathbf{T}_k$ and $\mathbf{V}_k$ parameters to be free. It is noted $[\mathbf{R}_k, \mathbf{T}_k, \mathbf{V}_k]$ and it corresponds to a standard heteroscedastic Gaussian mixture.

In the homoscedastic case $\boldsymbol{\Sigma}_k$ $(k = 1, \ldots, K)$ matrices are equal and so are $\mathbf{T}_k$ and $\mathbf{R}_k$ matrices since the decomposition (5) is unique. Then the homoscedastic model is denoted $[\mathbf{R}, \mathbf{T}, \mathbf{V}_k]$.

Table 1, where $[\bullet, a_k\mathbf{T}, \bullet]$ denotes a model of isotropically transformed standard deviations, indicates the parameter dimension of each model within the RTV family.

model	dimension.
$[\mathbf{R}_k, \mathbf{T}_k, \mathbf{V}_k]$ (general)	$Kd + Kd(d+1)/2$
$[\mathbf{R}_k, \mathbf{T}_k, \mathbf{V}]$	$d + Kd(d+1)/2$
$[\mathbf{R}_k, a_k\mathbf{T}, \mathbf{V}_k]$	$Kd + d + (K-1) + Kd(d-1)/2$
$[\mathbf{R}_k, a_k\mathbf{T}, \mathbf{V}]$	$2d + (K-1) + Kd(d-1)/2$
$[\mathbf{R}_k, \mathbf{T}, \mathbf{V}_k]$	$Kd + d + Kd(d-1)/2$
$[\mathbf{R}_k, \mathbf{T}, \mathbf{V}]$	$2d + Kd(d-1)/2$
$[\mathbf{R}, \mathbf{T}_k, \mathbf{V}_k]$	$2Kd + d(d-1)/2$
$[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$	$Kd + d(d+1)/2$
$[\mathbf{R}, a_k\mathbf{T}, \mathbf{V}_k]$	$Kd + (K-1) + d(d+1)/2$
$[\mathbf{R}, a_k\mathbf{T}, \mathbf{V}]$	$(K-1) + d(d+3)/2$
$[\mathbf{R}, \mathbf{T}, \mathbf{V}_k]$ (homoscedastic)	$Kd + d(d+1)/2$

Table 1: *Dimension of the Gaussian parameter of the RTV models.*

3.2 Graphical representations

In this section we propose a specific representation of Gaussian mixtures, which enables to highlight the homogeneity (or the heterogeneity) of the statistical parameters involved by the RTV models.

We refer now to Figure 3. The Gaussian parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of some normal random vector $\boldsymbol{Y}$ in $\mathbb{R}^2$, can be represented by: $\Gamma(\rho, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{\boldsymbol{x} \in \mathbb{R}^2; (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \rho\}$. The latter is an ellipsis the points of which are at a distance ρ from $\boldsymbol{\mu}$, according to the Mahalanobis metric $\boldsymbol{\Sigma}^{-1}$. The smallest rectangle containing Γ, plotted in dashed line, indicates the dispersion of $\boldsymbol{Y}$

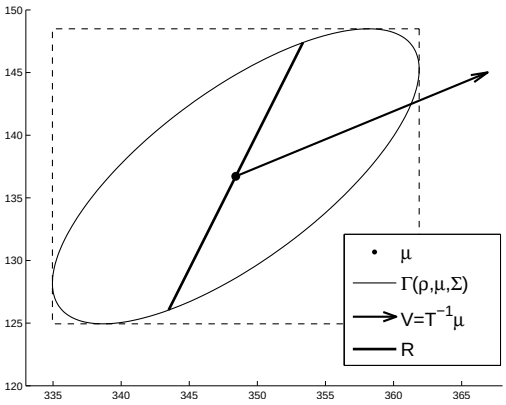
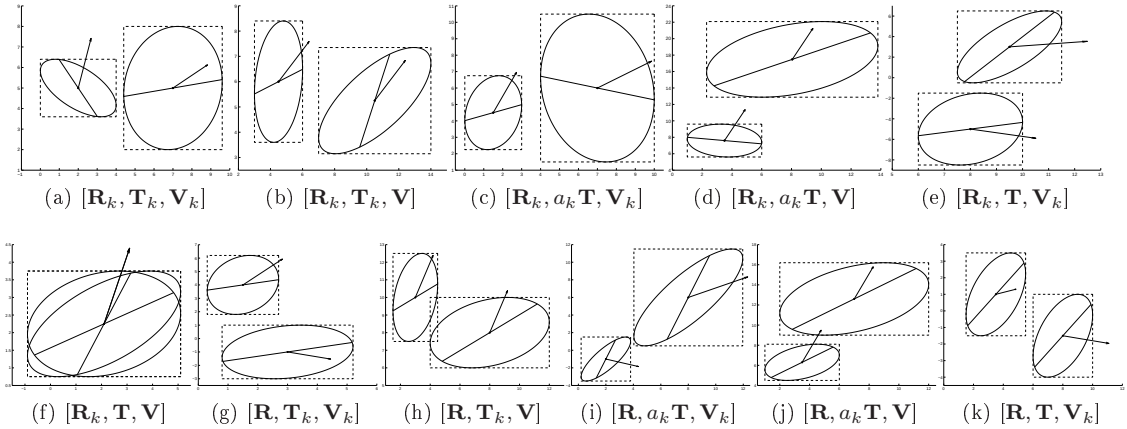


Figure 3: *Representation of the first-order-standardized-moments (arrow), of the standard deviations (dashed rectangle) and of the correlation (solid segment), for a random vector in* $\mathbb{R}^2$.

variables, and allows (possibly) to compare the dispersion of the corresponding variables of several Gaussian random vectors. The correlation of $\boldsymbol{Y}$ variables is represented by a solid segment centered in $\boldsymbol{\mu}$. The angle between the latter and the horizontal is proportional to the correlation of the variables, and the coefficient of proportionality is $\pi/2$. Thus, $\boldsymbol{Y}$ variables are even more close to independence (resp. even more correlated) as the solid segment is close to horizontal (resp. to vertical). The solid arrow with origin $\boldsymbol{\mu}$ represents $\boldsymbol{Y}$ first-order-standardized-moments that is $\boldsymbol{V} = \boldsymbol{T}^{-1}\boldsymbol{\mu}$ where $\boldsymbol{T}$ is the diagonal matrix of $\boldsymbol{Y}$ standard deviations. The drawn vector is $\boldsymbol{V}/\gamma$ where $\gamma = \|\boldsymbol{V}\|_2 \Big/ \left(\rho\sqrt{\sum_{i=1}^{2}\boldsymbol{T}(i,i)^2/2}\right)$ is a coefficient of graphical normalization by which the dimensions of the solid arrow and the dashed rectangle are close.

Then Figures 4a to 4k display the eleven models of the RTV family (for $d = 2$ and $K = 2$). The dashed rectangles enable to compare the conditional standard deviations, the solid segments, the correlation of the variables and the arrows, the first-order-standardized-moments. The vectors representing the latter are $\boldsymbol{V}_k/\gamma$ where the graphical normalization coefficient is now $\gamma = \left(\sum_{k=1}^{K}\|\boldsymbol{V}_k\|_2/K\right) \Big/ \left(\rho\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{2}\boldsymbol{T}_k(i,i)^2/(2K)}\right)$.



(a) $[\mathbf{R}_k, \mathbf{T}_k, \mathbf{V}_k]$ (b) $[\mathbf{R}_k, \mathbf{T}_k, \mathbf{V}]$ (c) $[\mathbf{R}_k, a_k\mathbf{T}, \mathbf{V}_k]$ (d) $[\mathbf{R}_k, a_k\mathbf{T}, \mathbf{V}]$ (e) $[\mathbf{R}_k, \mathbf{T}, \mathbf{V}_k]$

(f) $[\mathbf{R}_k, \mathbf{T}, \mathbf{V}]$ (g) $[\mathbf{R}, \mathbf{T}_k, \mathbf{V}_k]$ (h) $[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$ (i) $[\mathbf{R}, a_k\mathbf{T}, \mathbf{V}_k]$ (j) $[\mathbf{R}, a_k\mathbf{T}, \mathbf{V}]$ (k) $[\mathbf{R}, \mathbf{T}, \mathbf{V}_k]$

Figure 4: *Eleven Gaussian mixtures based on the parsimony of statistical parameters.*

4 PROPERTIES OF THE NEW MODELS

The proofs of the four following properties are reported in Appendix.

4.1 Faithful representations in low dimension

Property 1 shows that the constraints characterizing any RTV model remain in every canonical plane. This property ensures that the illustrations of Figure 4, are appropriate to represent the RTV models in any canonical plane whatever is $d \geq 2$. Reciprocally Property 2 establishes that a Gaussian mixture belongs to the RTV family if the same combination of RTV constraints holds in every canonical plane.

Property 1 (Stability of each RTV model by projection into any canonical plane)**.** $\mathbf{X}$ *is a random vector in* $\mathbb{R}^d$ *(*$d \geq 2$*) distributed according to a RTV model and* $\tilde{\mathbf{X}}$ *is a random vector in* $\mathbb{R}^2$*, its components being two distinct variables of* $\mathbf{X}$*. Then* $\tilde{\mathbf{X}}$ *is distributed as a 2-dimensional RTV model with identical constraints as* $\mathbf{X}$ *in* $\mathbb{R}^d$*.*

Figure 5 illustrates the latter property about the model $[\mathbf{R}, \mathbf{T}_k, \mathbf{V}_k]$. The ellipsoÃ¯ds in $\mathbb{R}^3$ represent two Gaussian random vectors with homogeneous correlations, free standard deviations and free first-order-standardized-moments. This RTV constraint combination produces identical slopes for each couple of solid segments within each canonical plane.

Property 2 (Characterization of the RTV models by the dimension 2)**.** $\mathbf{X}$ *is a random vector in* $\mathbb{R}^d$ *(*$d \geq 2$*) which projections into the canonical planes are submitted to a same combination of RTV constraints. Then* $\mathbf{X}$ *is distributed in* $\mathbb{R}^d$ *as a RTV model submitted to the same RTV constraints as the projections in* $\mathbb{R}^2$*.*

Then Figure 5 is typical of the model $[\mathbf{R}, \mathbf{T}_k, \mathbf{V}_k]$: a couple of solid segments within one canonical plane at less, would have distinct slopes if the two Gaussian random vectors in $\mathbb{R}^3$ were not homogeneously correlated.

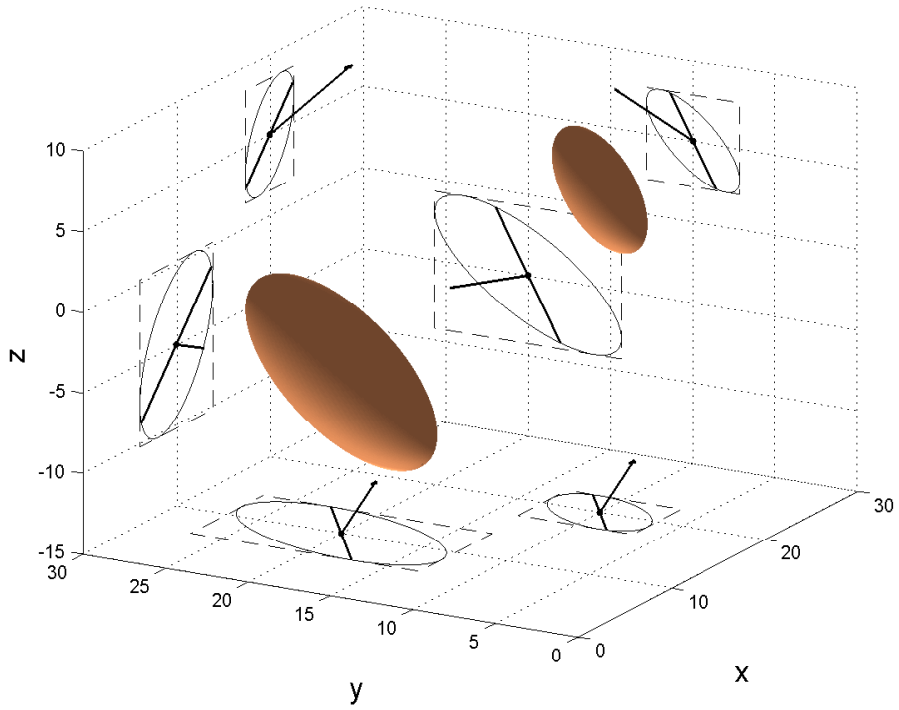


Figure 5: *Sustainability of the RTV structure by projection into the canonical planes: illustration on the model* $[\mathbf{R}, \mathbf{T}_k, \mathbf{V}_k]$.

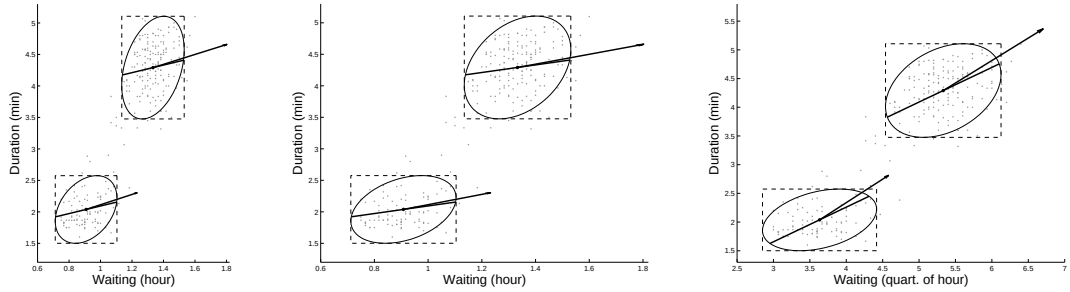
4.2 Scale invariance

In this section we show first that the constraints characterizing each RTV model still remain when the measurement units are changed (Property 3). Secondly we establish that changing the units has not even any effect on the model selection when the latter is based on classical criteria like *AIC* (Akaike 1974), *BIC* (Schwarz 1978) or *ICL* (Biernacki, Celeux, and Govaert 2000) (Property 4).

Property 3 (Stability of each RTV model by a linear transformation)**.** $\mathbf{X}$ *is a random vector in* $\mathbb{R}^d$*, distributed according to some RTV model.* $\mathbf{D} \in \mathbb{R}^{d \times d}$ *is a diagonal matrix, definite and positive. Then* $\mathbf{DX}$ *is distributed according to the same RTV model as* $\mathbf{X}$*.*

The illustrations provided by Figure 4 (defined in Subsection 3.2) are consistent with Property 3. In particular they allow to represent any RTV model whatever are (i) the axis scales and (ii) the graphical measurement units of the data. Figure 6a represents a model of homogeneous correlations (and free other parameters) inferred on the Old Faithful data when Waiting and Duration are measured respectively in hour and minute (see data description in Subsection 6.1). Figure 6b shows that–unlike the orientations (see Figure 2)–the correlations still appear as homogeneous when an axis is rescaled (the solid segments still have the same slope in Figure 6b). In addition the correlations appear as homogeneous even if the Waiting variable is represented

in quarter of hour instead of hour (see Figure 6c).



(a) *Same correlations.* (b) *Rescaling x-axis.* (c) *Changing the Waiting unit.*

Figure 6: *Sustainability of the RTV structure for any axis scale ratio and for any graphical measurement units of the data.*

Property 4 (Invariance of the model selection to the unit modification)**.** *The choice of some RTV model thanks to AIC, BIC or ICL, does not depend on the measurement units of the data.*

Then, as an immediate consequence, a model selected in the RTV family and the associated partition are insensible to the reduction of the variables.

5 PARAMETER ESTIMATION

5.1 Overview

The algorithm which enables to estimate $\boldsymbol{\psi}$ by maximizing its likelihood depends on the considered model. An EM algorithm (Dempster, Laird, and Rubin 1977) can be implemented in the standard homoscedastic and heteroscedastic cases (see McLachlan and Peel 2000). For the nine remaining RTV models, a Generalized EM algorithm (Dempster et al. 1977) is required. The latter increases iteratively $\boldsymbol{\psi}$ likelihood by alternating the two following steps:

- E-step. Compute the conditional probabilities t_i^k related to the current value of $\boldsymbol{\psi}$, according to (2).

- GM-step. Increase–instead of maximize in an EM algorithm–the expected log-likelihood of $\boldsymbol{\psi}$ parameter:

$$\sum_{i=1}^{n}\sum_{k=1}^{K} t_i^k\{\log \pi_k + \log \Phi_d\left(\mathbf{x}_i; \mathbf{T}_k\mathbf{V}_k, \mathbf{T}_k\mathbf{R}_k\mathbf{T}_k\right)\}. \tag{6}$$

Remind that $\mathbf{V}_k = \mathbf{T}_k^{-1}\boldsymbol{\mu}_k$. As (6) is additively separable with respect to $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ on one hand and to $\boldsymbol{\nu} = (\mathbf{V}_1, \ldots, \mathbf{V}_K)$, $\boldsymbol{\tau} = (\mathbf{T}_1, \ldots, \mathbf{T}_K)$ and $\boldsymbol{\rho} = (\mathbf{R}_1, \ldots, \mathbf{R}_K)$ on the other hand, the GM step is decomposed itself into:

- Step GM-1. Estimation of the mixing-proportions when they are supposed to be free, according to the classical formula: $\hat{\pi}_k = \hat{n}_k/n$ where $\hat{n}_k = \sum_{i=1}^n t_i^k$.

- Step GM-2. Increase of (6) with respect alternately to the components $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ of $\boldsymbol{\psi}$.

The second step (GM-2) amounts to decrease the criterion:

$$\sum_{k=1}^{K}\sum_{i=1}^{n} t_i^k \{2\log|\mathbf{T}_k| + \log|\mathbf{R}_k| + (\mathbf{T}_k^{-1}\mathbf{x}_i - \mathbf{V}_k)'\mathbf{R}_k^{-1}(\mathbf{T}_k^{-1}\mathbf{x}_i - \mathbf{V}_k)\}. \tag{7}$$

The following subsection details the three iterations constituting the GM-2 step.

5.2 Detail of the GM-2 step

Estimation of $\boldsymbol{\nu}$ $\bar{\mathbf{x}}_k = (1/\hat{n}_k)\sum_{i=1}^n t_i^k\,\mathbf{x}_i$ is the expected empirical mean in class k. Considering $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ as fixed components of $\boldsymbol{\psi}$ parameter, (7) reaches its minimum for $\mathbf{V}_k = \mathbf{T}_k^{-1}\bar{\mathbf{x}}_k$ when the conditional first-order-standardized-moments are supposed to be free, and for

$$\mathbf{V} = \left(\sum_{k=1}^{K}\hat{n}_k\mathbf{R}_k^{-1}\right)^{-1}\left(\sum_{k=1}^{K}\hat{n}_k\mathbf{R}_k^{-1}\mathbf{T}_k^{-1}\bar{\mathbf{x}}_k\right) \tag{8}$$

when the first-order-standardized-moments are supposed to be homogeneous.

Estimation of $\boldsymbol{\tau}$ When the components $\boldsymbol{\nu}$ and $\boldsymbol{\rho}$ are fixed, three cases must be considered depending on the constraint set on the standard deviations (free, isotropically transformed or homogeneous).

- When $\mathbf{T}_k$ matrices are homogeneous ($\mathbf{T}_k = \mathbf{T}$), minimizing (7) amounts to determine the minimum $\boldsymbol{x}_0$ of:

$$-2\log|\mathrm{diag}\,\boldsymbol{x}| - 2\mathbf{L}\boldsymbol{x} + \boldsymbol{x}'\mathbf{Q}\boldsymbol{x} \tag{9}$$

where $\boldsymbol{x} \in (\mathbb{R}_*^+)^d$, diag $\boldsymbol{x}$ denotes the diagonal matrix constituted of $\boldsymbol{x}$ components,

$$\mathbf{Q} = (1/n)\sum_{k=1}^{K}\sum_{i=1}^{n} t_i^k(\mathrm{diag}\,\mathbf{x}_i)\mathbf{R}_k^{-1}(\mathrm{diag}\,\mathbf{x}_i) \tag{10}$$

and

$$\mathbf{L} = \sum_{k=1}^{K}(\hat{n}_k/n)\mathbf{V}_k'\mathbf{R}_k^{-1}(\mathrm{diag}\,\bar{\mathbf{x}}_k). \tag{11}$$

As $\mathbf{Q}$ is symmetric, definite, positive, (9) is convex with respect to $\boldsymbol{x}$ and one can get close to $\boldsymbol{x}_0$ thanks to any convex optimization algorithm. Then the researched matrix is $\mathbf{T} = (\mathrm{diag}\,\boldsymbol{x}_0)^{-1}$.

- When $\mathbf{T}_k$ matrices are mutually and isotropically transformed then for each k: $\mathbf{T}_k = a_{1,k}\mathbf{T}_1$ ($a_{1,k} > 0$). We note $a_k = a_{1,k}$ ($k = 1, \ldots, K$). Since the minimum of (7) with respect to $\boldsymbol{\tau}$ is not explicit, one decreases (7) by alternated minimizations with respect to a_k coefficients and to $\mathbf{T}_1$ matrix. For fixed a_k ($k = 1, \ldots, K$), (7) is convex with respect to $\mathbf{T}_1^{-1}$; then one determines the $\mathbf{T}_1$ value corresponding to the minimum of (7). For some fixed $\mathbf{T}_1$ matrix, (7) is minimal when ($k = 2, \ldots, K$ and $\mathbf{1}_d = (1, \ldots, 1)' \in \mathbb{R}^d$):

$$a_k = \frac{-\mathbf{L}_k(\mathbf{T}_1^{-1}\mathbf{1}_d) + \sqrt{\left[\mathbf{L}_k(\mathbf{T}_1^{-1}\mathbf{1}_d)\right]^2 + 4d\left[(\mathbf{T}_1^{-1}\mathbf{1}_d)'\mathbf{R}_k^{-1}(\mathbf{T}_1^{-1}\mathbf{1}_d)\right]}}{2d}. \tag{12}$$

- When $\mathbf{T}_k$ matrices are free, the minimum of (7) with respect to $\boldsymbol{\tau}$ is obtained by determining sequentially (and independently) the k minima of (9) corresponding to the parameters:

$$\mathbf{Q}_k = (1/\hat{n}_k)\sum_{i=1}^{n} t_i^k(\text{diag } \mathbf{x}_i)\mathbf{R}_k^{-1}(\text{diag } \mathbf{x}_i) \tag{13}$$

and

$$\mathbf{L}_k = \mathbf{V}_k'\mathbf{R}_k^{-1}(\text{diag } \bar{\mathbf{x}}_k). \tag{14}$$

One can get close to each minimum $\boldsymbol{x}_k$ thanks to any convex optimization algorithm since $\mathbf{Q}_k$ matrix is symmetric, definite and positive. Then the researched matrices $\mathbf{T}_k$ are given by: $\mathbf{T}_k = (\text{diag } \boldsymbol{x}_k)^{-1}$.

Estimation of $\boldsymbol{\rho}$ When the components $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ of $\boldsymbol{\psi}$ are fixed, the criterion (7) to minimize becomes

$$\sum_{k=1}^{K} \hat{n}_k \left[\log|\mathbf{R}_k| + \text{tr}\left(\mathbf{W}_k\mathbf{R}_k^{-1}\right)\right] \tag{15}$$

where $\mathbf{W}_k = (1/\hat{n}_k)\sum_{i=1}^{n} t_i^k(\mathbf{T}_k^{-1}\mathbf{x}_i - \mathbf{V}_k)(\mathbf{T}_k^{-1}\mathbf{x}_i - \mathbf{V}_k)'$, if the conditional correlations are supposed to be free, and

$$\log|\mathbf{R}| + \text{tr}\left(\mathbf{W}\mathbf{R}^{-1}\right) \tag{16}$$

where $\mathbf{W} = \sum_{k=1}^{K}(\hat{n}_k/n)\mathbf{W}_k$, if the conditional correlations are supposed homogeneous. (16) can be decreased alternately with respect to each correlation of $\mathbf{R}$. Indeed, fixing all the correlations of $\mathbf{R}$ except one of them, (16) has limit $+\infty$ at -1, $+\infty$ at 1, and two local minima at most between -1 and 1 (see Lourme 2011, pp. 83–86). In case of free conditional correlations, lessening (16) amounts to decrease independently the terms $\log|\mathbf{R}_k| + \text{tr}\left(\mathbf{W}_k\mathbf{R}_k^{-1}\right)$ ($k = 1, \ldots, K$) whith respect to the coefficients of $\mathbf{R}_k$ matrices.

6 EXPERIMENTS ON REAL DATA

6.1 Clustering of Old Faithful eruptions

We consider $n = 272$ eruptions of the famous Old Faithful geyser, described by two variables ($d = 2$): Duration (of an eruption) and Waiting (to the next eruption) both measured in minutes. This sample from Venables and Ripley 2002 has been subject of many clustering studies

(see Atkinson and Riani 2007 for an example) and the most widespread structure of Old Faithful eruptions in the literature consists of two clusters (often interpreted as short and long eruptions). On the other hand we have observed that in whatever family (RTV or geometrical) a model is selected, ICL infers two classes of eruptions whereas BIC leads to three clusters. Then we set $K = 2$ in the following and we base the model selection on ICL.

Each model of both families has been inferred on the previous geyser data and Table 2a displays the four best models of each family (according to ICL).

family	rank	model	ICL	model	ICL	model	ICL
geometrical	1	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	1160.3	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	**2272.4**	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}\boldsymbol{S}_k']$	414.57
	2	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k']$	1161.4	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k']$	2275.0	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}_k\boldsymbol{S}_k']$	415.55
	3	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}']$	1161.7	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$	2275.1	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}']$	415.89
	4	$(\pi_k)[\lambda_k \boldsymbol{S}_k\boldsymbol{\Lambda}\boldsymbol{S}_k']$	1162.9	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}']$	2275.4	$(\pi_k)[\lambda \boldsymbol{S}_k\boldsymbol{\Lambda}\boldsymbol{S}_k']$	417.02
RTV	1	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	**1158.8**	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	2272.5	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$	**412.99**
	2	$(\pi_k)[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	1161.4	$(\pi_k)[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	2275.0	$(\pi_k)[\boldsymbol{R}_k, \boldsymbol{T}_k, \boldsymbol{V}_k]$	415.55
	3	$(\pi_k)[\boldsymbol{R}, a_k\boldsymbol{T}, \boldsymbol{V}_k]$	1161.7	$(\pi_k)[\boldsymbol{R}, a_k\boldsymbol{T}, \boldsymbol{V}_k]$	2275.4	$(\pi_k)[\boldsymbol{R}, a_k\boldsymbol{T}, \boldsymbol{V}_k]$	415.89
	4	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	1163.4	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	2277.0	$(\pi_k)[\boldsymbol{R}, \boldsymbol{T}, \boldsymbol{V}_k]$	417.55
		(a) *min*×*min (original units)*		(b) *sec*×*min*		(c) *reduced*×*reduced*	

Table 2: *The four best models within each family (RTV and geometrical), inferred on the Old Faithful data ($K = 2$) when Duration × Waiting measurement units vary.*

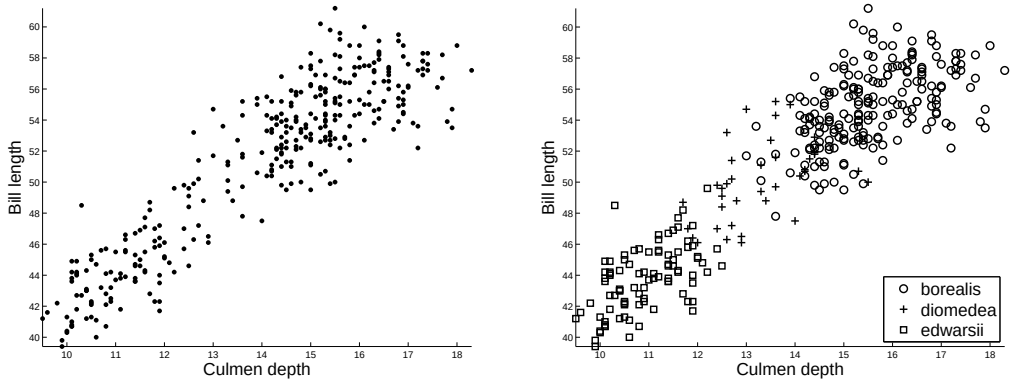
The best RTV model $(\pi_k)[\boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{V}_k]$ ($ICL = 1158.8$) surpasses the best geometrical one $(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}_k\boldsymbol{S}']$ ($ICL = 1160.3$)–π_k in parentheses indicates free mixing proportions for both models–and provides also a more convincing interpretation of the conditional data properties. Indeed it assumes that Duration and Waiting are identically correlated among short and long eruptions; according to the best geometrical model, the eruption classes share an identical orientation which does not have so much sense for a geologist.

Now let us consider the effect of transforming the Duration (and this variable only) from minutes into seconds, on the ICL-associated rank of the models within their respective family. Table 2b confirms that the rank of each RTV model is not altered by modifying the measurement units whereas the ranks of some geometrical models are changed. The rank invariance of the RTV models rests on that all their ICL values differ by $272 \log 60$ from Table 2a to Table 2b. This could be expected from (23) since 272 is the sample size and 60 is the coefficient that transforms minutes into seconds.

The ICL-rank of the RTV models keeps unchanged even by reducing both variables Duration and Waiting as shown by Table 2c. So, reducing both variables does not affect the model selected in the RTV family whereas it modifies the model chosen in the geometrical one (see Tables 2a and 2c). Let us notice that for each RTV model, the ICL values in Tables 2a and 2c differ by $272 \log(a.b)$ where $a \approx 13.595$ and $b \approx 1.141$ are the respective standard deviations of Duration and Waiting. (This observation was predictable from (23).)

6.2 Clustering of seabirds

We consider $n = 336$ Cory's Shearwaters (which are seabirds from the species *Calonectris diomedea*), described by five morphological variables ($d = 5$): Culmen depth, Bill length, etc. (see Figure 7a). These birds studied by Thibault, Bretagnolle, and Rabouam (1997), are divided into three subspecies: *borealis*, *diomedea* and *edwardsii* (see Figure 7b).



(a) *The data (a sample of Cory's Shearwaters).* (b) *The researched partition (the bird subspecies).*

Figure 7: *Cory's Shearwaters from three subspecies.*

For each value of K from 1 to 5 all the models of both families (RTV and geometric) are inferred on the Shearwaters data keeping the original morphometric measurement units provided by the ornithologists. Table 3 displays the best BIC value obtained within each family. Here the

K	1	2	3	4	5
RTV models	4472.0	4356.6	**4335.2**	4347.8	4370.1
geometric models	4472.0	4362.5	4344.2	**4341.7**	4355.8

Table 3: *Best BIC values obtained by the models of both families (RTV and geometric) on Shearwaters data, for a variable cluster number (K).*

RTV family enables clearly to retrieve three bird clusters contrarily to the geometrical models which find four groups of Shearwaters.

The overall best model ($BIC = 4335.2$), obtained in the RTV family for $K = 3$ groups, is $(\pi_k)[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$. Table 4 shows that the associated partition is very close to the seabird subspecies since the error rate obtained by comparing both partitions is 2.68%. For the same cluster number, the best geometric model provides a worth BIC value (4344.2) and a worth error rate also (2.98%).

So the model $(\pi_k)[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$ is not only better than every geometrical model according to BIC but this model enables also to distinguish *borealis* from *diomedea* or *edwardsii* better than the best geometrical model would do.

family	best model	BIC	error rate
RTV	$(\pi_k)[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$	**4335.2**	**2.68%**
geometric	$(\pi_k)[\lambda_k \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}']$	4344.2	2.98%

Table 4: *Error rate (obtained by comparing the estimated partition to the bird subspecies) and BIC value of the best model whithin each family for $K = 3$ groups.*

According to this model the correlations and the coefficients of variation of the five biological variables (Bill length, Culmen depth, Tarsus, Wing and Tail) are homogeneous through the subspecies, whereas the conditional standard deviations differ. These characteristic features of the selected model are highlighted by Figure 8: whatever is the canonical plane in which the data and the inferred model are projected, the arrows are the same and the solid segments have equal slopes.

But the retained model $(\pi_k)[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$ is not only better than the best geometrical model for BIC and for the associated error rate. This model allows also a dynamical interpretation of the conditional distributions: every shearwater subspecies derives stochastically from a common reference population modelized by a Gaussian vector $\mathbf{X}_0$ the center of which is $\mathbf{V}$ (the vector composed of the homogenous first-order-standardized-moments) and the covariance matrix of which is $\mathbf{R}$ (the matrix composed of the homogeneous correlations of the variables). Indeed noting $(\mathbf{X}|Z_k = 1)$ $(k = 1, 2, 3)$ the Gaussian vectors modelizing respectively *borealis*, *diomedea* and *edwardsii* populations then for all k: $(\mathbf{X}|Z_k = 1) \stackrel{\mathscr{D}}{=} \mathbf{T}_k\mathbf{X}_0$, where $\mathbf{T}_k$ denotes the standard deviation matrix within the subspecies k ($\bullet \stackrel{\mathscr{D}}{=} \bullet$ indicates the equality of two random vectors in distribution).

Figure 9 represents both the selected model $(\pi_k)[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$, the estimated subspecies and the reference population from which *borealis*, *diomedea* and *edwardsii* arise stochastically according to the model. The hypothesis of some common origin for the three subspecies of Shearwaters rests on a mathematical interpretation of the retained model $(\pi_k)[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$. There would be a great interest to know whether this assumption makes sense for ornithologists.

7 CONCLUDING REMARKS

Using parsimonious mixtures to modelize heterogeneous data aims to find a compromise between the bias of the estimated model and its variability to the sampling fluctuations. But the choice of the parameters to which one applies the parsimony, cannot be reduced to this technical goal. The constraints set on the parameter must allow the parsimonious models to be both represented and interpreted.

So we have defined new multidimensional Gaussian mixtures the parsimony of which is about parameters of statistical interpretation: the standard deviation and the coefficient of variation of the variables, the correlation of the couples of variables.

These models, called RTV, own multiple properties which are determinant in their attractivity. The characteristic constraints of each RTV mixture (i) still remain in every canonical subspace and (ii) still remain by changing the measurement units. In addition the change of the

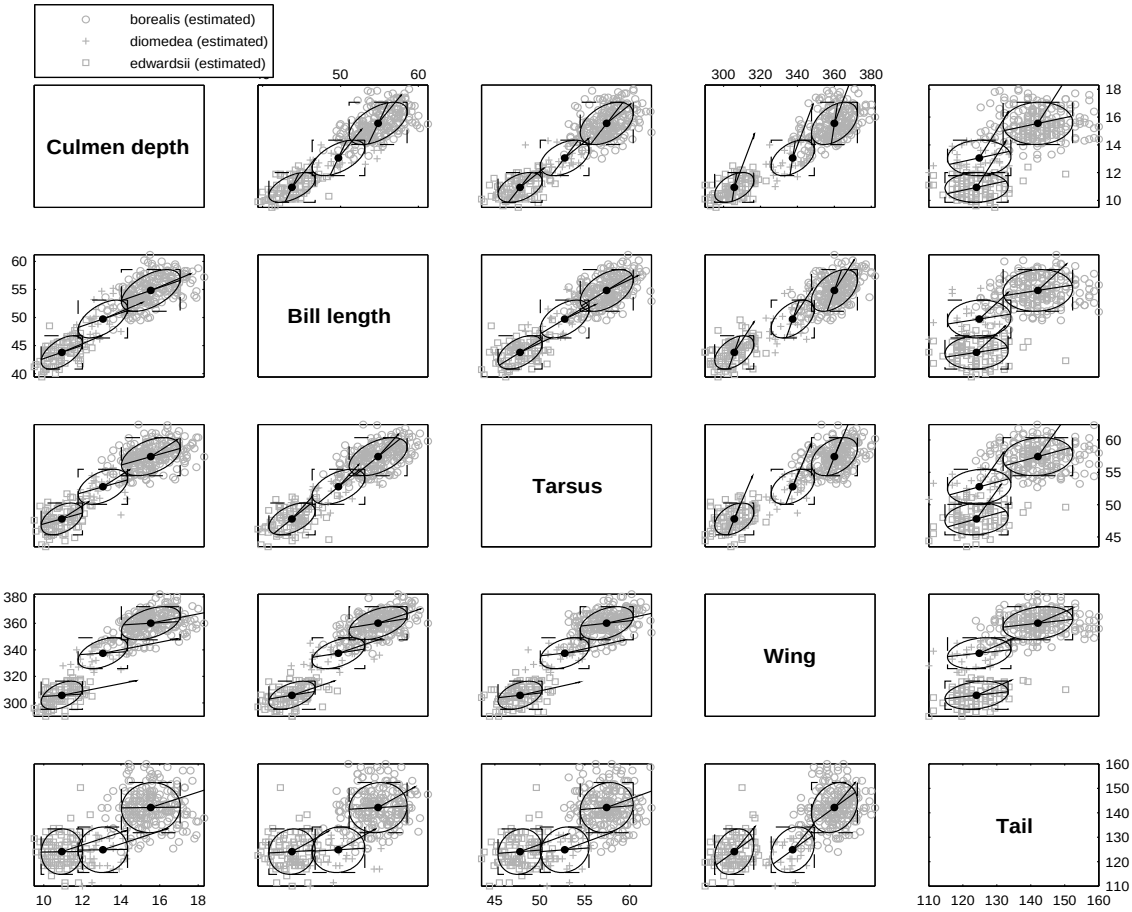


Figure 8: *The estimated partition of the Shearwaters and the selected model* $[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$ *in every canonical plane.*

units has no effect on the model selection based on some likelihood criteria like *AIC*, *BIC* or *ICL*. Typically the model selected in the RTV family is not modified by reducing the variables. In supervised inference situations, the misclassification error rate estimated by cross-validation is often used in order to select a model (see Govaert 2009, pp. 189–190). In these situations also, the retained RTV model will not depend on the units of the data. In particular the cross-validated error rate will lead to select the same RTV model on raw data or on reduced data.

The parsimonious Gausian mixtures of Celeux and Govaert (1995) based on a geometrical interpretation of the classes, do not own none of the previous properties. We have displayed two cases, one from geology and the other one from biology, where the geometrical models are supplanted by our statistical ones. In both examples, the RTV family provides a better model fit, a better classifier and a more relevant interpretation of the selected model.

Some of the RTV models allow for a dynamical interpretation of the inferred groups. Indeed we showed that a model where both correlations and coefficients of variation are homogeneous,

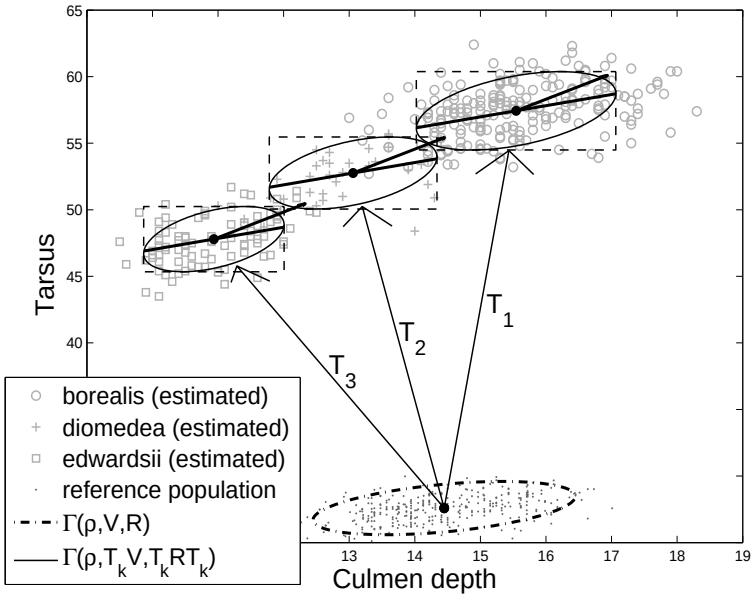


Figure 9: *Dynamical interpretation of the selected model* $(\pi_k)[\mathbf{R}, \mathbf{T}_k, \mathbf{V}]$*: the three subspecies of Shearwaters stochastically arise from a reference population.*

enables to assume that the conditional populations arise stochastically from a common reference population. The hypothesis of a common reference population is authorized by the five RTV models assuming homogeneous correlations (even if the reference population is not identifiable when the coefficients of variation are not themselves homogeneous). The dynamical interpretation of the clusters is a novelty in classification. Indeed traditional clustering methods, whether based on mixtures or on a geometric criterion optimization, all consider the classes from a static viewpoint. According to us it is pertinent to try to define in the future (as far as possible) new models which enable–like our RTV models–some dynamical interpretation of the classes. Such models would establish an interpretable link between the conditional populations no more only consisting on a static algebraic constraint on the parameters.

Diverse experiments led with the RTV models (of which Sections 6.1 and 6.2 are only two examples) show a predominance of the RTV models assuming homogeneous correlations. The success of such models can be explained (in part) by the large number of conditional correlations in a multidimensional mixture, and by the cost of their estimate when they are supposed to be free. For example in dimension $d = 7$ and for $K = 3$ components, the parameter size of an heteroscedastic Gaussian mixture decreases by 40% when the conditional correlations are supposed to be homogeneous. However, the models assuming homogeneous correlations are recurrently better than the other RTV models in biology, but also better than the geometrical models. Then we invite the biologists who use the RTV models to consider from an expert viewpoint, the possibility for biological variables to be uniformly correlated in heterogeneous populations.

The statistical parameters concerned by the parsimony of the RTV models (correlations, standard deviations, coefficients of variation) are not specific to the Gaussian mixtures. One can extend the RTV models to any mixtures, the second-order-conditional-moments of which

are finite. For example one will be able to envisage parsimonious multivariate Student mixtures (McLachlan and Peel 2000) based on the decomposition (5) of the covariance matrices, as soon as the conditional degrees of freedom are (strictly) greater than 2.

APPENDIX: PROOFS OF PROPERTIES

Proof of Property 1. Let us note $\boldsymbol{\mu}_k$, $\mathbf{R}_k$, $\mathbf{T}_k$ the conditional parameters of $\mathbf{X}$.
There exists a matrix $\mathbf{P}$ in $\{0,1\}^{2\times d}$ having exactly one 1 per row, at most one 1 per column, and such that $\tilde{\mathbf{X}} = \mathbf{PX}$.
Each conditional vector $(\tilde{\mathbf{X}}|Z_k = 1)$ is Gaussian with center: $\tilde{\boldsymbol{\mu}}_k = \mathbf{P}\boldsymbol{\mu}_k$ and covariance matrix:

$$\tilde{\boldsymbol{\Sigma}}_k = \mathbf{PT}_k\mathbf{R}_k\mathbf{T}_k\mathbf{P}'. \tag{17}$$

But the latter can also be written:

$$\tilde{\boldsymbol{\Sigma}}_k = (\mathbf{PT}_k\mathbf{P}')(\mathbf{PR}_k\mathbf{P}')(\mathbf{PT}_k\mathbf{P}'). \tag{18}$$

The diagonal coefficients of $\tilde{\mathbf{R}}_k = \mathbf{PR}_k\mathbf{P}'$ are equal to 1 and the other coefficients are between -1 and 1; the matrix $\tilde{\mathbf{T}}_k = \mathbf{PT}_k\mathbf{P}'$ is diagonal. Then $\tilde{\mathbf{T}}_k$ is the standard deviation matrix of $(\tilde{\mathbf{X}}|Z_k = 1)$ and $\tilde{\mathbf{R}}_k$ is its correlation matrix.

Assuming that the conditional correlations of $\mathbf{X}$ are homogeneous ($\mathbf{R}_k = \mathbf{R}$) involves that those of $\tilde{\mathbf{X}}$ are also homogeneous ($\tilde{\mathbf{R}}_k = \mathbf{PRP}'$).

A constraint set on the standard deviations of $\mathbf{X}$ involves an identical constraint on the standard deviations of $\tilde{\mathbf{X}}$. If the conditional standard deviations of $\mathbf{X}$ are isotropically transformed: $\mathbf{T}_{k'} = a_{k,k'}\mathbf{T}_k$ ($a_{k,k'} \in \mathbb{R}_+^*$) (resp. are equal: $\mathbf{T}_k = \mathbf{T}$), so are also those of $\tilde{\mathbf{X}}$: $\tilde{\mathbf{T}}_{k'} = a_{k,k'}\tilde{\mathbf{T}}_k$ (resp. $\tilde{\mathbf{T}}_k = \mathbf{PTP}'$).

At last if the conditional first-order-standardized-moments of $\mathbf{X}$ are homogeneous: $\mathbf{T}_k^{-1}\boldsymbol{\mu}_k = \mathbf{V}$, so are also those of $\tilde{\mathbf{X}}$: $\tilde{\mathbf{T}}_k^{-1}\tilde{\boldsymbol{\mu}}_k = \mathbf{PV}$ (This equality results from: $(\mathbf{PT}_k\mathbf{P}')^{-1} = \mathbf{PT}_k^{-1}\mathbf{P}'$.). □

Proof of Property 2. This property rests on the following obvious argument. Two random vectors of $\mathbb{R}^d$ have homogeneous correlations when their projections onto any canonical plane have themselves homogeneous correlations. This argument can easily be extended to the constraints on standard deviations or on first-order-standardized-moments.

□

Proof of Property 3. Let us note $\boldsymbol{\mu}_k$, $\mathbf{R}_k$, $\mathbf{T}_k$ the conditional parameters of $\mathbf{X}$ and $\tilde{\boldsymbol{\mu}}_k$, $\tilde{\mathbf{R}}_k$, $\tilde{\mathbf{T}}_k$ those of $\mathbf{DX}$.
$\mathbf{X}$ and $\mathbf{DX}$ share the same conditional correlations: $\tilde{\mathbf{R}}_k = \mathbf{R}_k$ (resp. the same conditional first-order-standardized-moments: $\tilde{\mathbf{T}}_k^{-1}\tilde{\boldsymbol{\mu}}_k = {\mathbf{T}_k}^{-1}\boldsymbol{\mu}_k$). Assuming that the conditional correlations of $\mathbf{X}$ are homogeneous ($\mathbf{R}_k = \mathbf{R}$) involves that those of $\mathbf{DX}$ are homogeneous also ($\tilde{\mathbf{R}}_k = \mathbf{R}$). Similarly, assuming that $\mathbf{X}$ conditional first-order-standardized-moments are equal (${\mathbf{T}_k}^{-1}\boldsymbol{\mu}_k = \mathbf{V}$) involves that those of $\mathbf{DX}$ are equal also ($\tilde{\mathbf{T}}_k^{-1}\tilde{\boldsymbol{\mu}}_k = \mathbf{V}$).

On the other hand, since $\tilde{\mathbf{T}}_k = \mathbf{DT}_k$, if the conditional standard deviations of $\mathbf{X}$ are isotropically transformed ($\mathbf{T}_{k'} = a_{k,k'}\mathbf{T}_k$) (resp. are equal ($\mathbf{T}_k = \mathbf{T}$)) through the classes, then those of $\mathbf{DX}$ are also isotropically transformed (resp. are also equal).

□

Proof of Property 4. $\mathbf{x} = \{\mathbf{x}_i; i = 1, \ldots, n\}$ is a sample in $\mathbb{R}^d$ and $\boldsymbol{D}\mathbf{x} = \{\boldsymbol{D}\mathbf{x}_i; i = 1, \ldots, n\}$, where $\boldsymbol{D} \in \mathbb{R}^{d \times d}$ is a diagonal matrix, definite and positive (the matrix of the measurement units modification).

Let us consider some RTV model the parameter of which is $\boldsymbol{\psi}$.
Property 3 ensures that modifying the measurement units of the data amounts to reparameterize the model at hand. Then:

$$\ell(\hat{\boldsymbol{\psi}}; \mathbf{x}) - \ell(\hat{\boldsymbol{\psi}}; \boldsymbol{D}\mathbf{x}) = n \log |\boldsymbol{D}|, \tag{19}$$

where $\ell(\hat{\boldsymbol{\psi}}; \mathbf{x})$ and $\ell(\hat{\boldsymbol{\psi}}; \boldsymbol{D}\mathbf{x})$ denote the maximized log-likelihood of $\boldsymbol{\psi}$ parameter, computed respectively on the data $\mathbf{x}$ and $\boldsymbol{D}\mathbf{x}$. The difference between both maximized log-likelihoods depends on the sample size (n), on the volume of the measurement units transformation ($|\mathbf{D}|$), but not on the RTV model at hand.
So, for all model $\mathscr{M}$ in the RTV family:

$$BIC(\mathscr{M}; \mathbf{x}) - BIC(\mathscr{M}; \boldsymbol{D}\mathbf{x}) = n \log |\boldsymbol{D}|. \tag{20}$$

Changing the measurement units according to $\boldsymbol{D}$ does translate all BIC values of the RTV models from a common term $n \log |\boldsymbol{D}|$. But such a modification does not change the rank (according to BIC) of some model whithin the RTV family.

Equality (20) holds also for the criterion AIC (Akaike 1974) and the proof is similar to the previous one written for BIC.

Noting $\hat{\boldsymbol{\psi}}(\mathbf{x})$ the parameter of a model inferred (by Maximum Likelihood) on $\mathbf{x}$ data, then for anyone of the RTV models, the estimators $\hat{\boldsymbol{\psi}}(\mathbf{x})$ and $\hat{\boldsymbol{\psi}}(\mathbf{D}\mathbf{x})$ are linked by the following relations:

$$\hat{\boldsymbol{\mu}}_k(\mathbf{D}\mathbf{x}) = \mathbf{D}\hat{\boldsymbol{\mu}}_k(\mathbf{x}) \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_k(\mathbf{D}\mathbf{x}) = \mathbf{D}\hat{\boldsymbol{\Sigma}}_k(\mathbf{x})\mathbf{D} \ ; \ k = 1, \ldots, K. \tag{21}$$

Then we deduce that for all $i \in \{1, \ldots, n\}$ and all $k \in \{1, \ldots, K\}$:

$$\Phi_d(\mathbf{D}\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k(\mathbf{D}\mathbf{x}), \hat{\boldsymbol{\Sigma}}_k(\mathbf{D}\mathbf{x})) = |\mathbf{D}|^{-1} \Phi_d(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_k(\mathbf{x}), \hat{\boldsymbol{\Sigma}}_k(\mathbf{x})), \tag{22}$$

where $\Phi_d(\bullet; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the d-dimensional normal density with center $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then the parameters $\hat{\boldsymbol{\psi}}(\mathbf{x})$ and $\hat{\boldsymbol{\psi}}(\mathbf{D}\mathbf{x})$ lead to identical conditional probabilities (2), on the data $\mathbf{x}$ and $\mathbf{D}\mathbf{x}$ respectively. So the entropy term by which ICL differ from BIC is the same for $\hat{\boldsymbol{\psi}}(\mathbf{x})$ and $\hat{\boldsymbol{\psi}}(\mathbf{D}\mathbf{x})$. This is why (20) can be extended to the ICL criterion:

$$ICL(\mathscr{M}; \mathbf{x}) - ICL(\mathscr{M}; \boldsymbol{D}\mathbf{x}) = n \log |\boldsymbol{D}|. \tag{23}$$

So the choice of some model within the RTV family, based on whatever criterion BIC, AIC or ICL, does not depend on the measurement units.

□

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

[2] Atkinson, A., and Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis*, 52(1):272–285.

[3] Banfield, J.D., and Raftery, A.E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49:803–821.

[4] Baudry, J.P. (2009). *Sélection de Modèle pour la Classification Non Spervisée. Choix du Nombre de Classes.* Thèse de doctorat, Université Paris-Sud 11.

[5] Biernacki, C. (1997). *Choix de modèles en classification.* Thèse de doctorat, Université Technologie de Compiègne.

[6] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725.

[7] Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics and Data Analysis*, 51(2):587–600.

[8] Bouveyron, C. (2006). *Modélisation et classification des données de grande dimension : application à l'analyse d'images.* Thèse de doctorat, Université Grenoble 1.

[9] Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52(1):502–519.

[10] Celeux, G., and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793.

[11] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society series B*, 39(1):1–38.

[12] Govaert, G. (2009). *Data Analysis.* Wiley.

[13] Greselin, F., Ingrassia, S., and Punzo, A. (2011). Assessing the pattern of covariance matrices via an augmentation multiple testing procedure. *Statistical Methods & Applications*, 20:141–170.

[14] Lourme, A. (2011). *Contribution Ã la Classification par Modèles de Mélange et Classification Simultanée d'Echantillons d'Origines Multiples.* Thèse de doctorat, Université Lille 1.

[15] Maugis, C., Martin-Magniette, M.L., Tamby, J.P., Renou, J.P., Lecharny, A., Aubourg, S., and Celeux, G. (2009). Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins. *MODULAD*, 40.

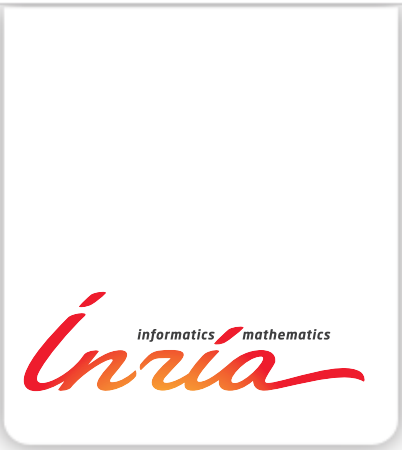
[16] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley.

[17] Michel, B. (2008). *Modélisation de la production d'hydrocarbures dans un bassin pétrolier.* Thèse de doctorat, Université Paris 11.

[18] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

[19] Thibault, J.C., Bretagnolle, V., and Rabouam, C. (1997). Cory's shearwater calonectris diomedea. *Birds of Western Paleartic Update*, 1:75–98.

[20] Venables, W.N., and Ripley, B.D. (2002). *Modern Applied Statistics with S. Fourth Edition.* Springer, New York.



RESEARCH CENTRE
LILLE – NORD EUROPE

Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399